Background: Operation of a Point-Contact Transistor

This document outlines the operation of a point-contact transistor (the first type of transistor). This information is to be used in creating a video on the origins of the integrated circuit. The goal is to a) describe the operation *this* transistor as a *current* amplifier, b) to show how it can be used as a switch (saturation vs cutoff), and c) how that switch can be used to create logic gates (AND, OR, and NOT).

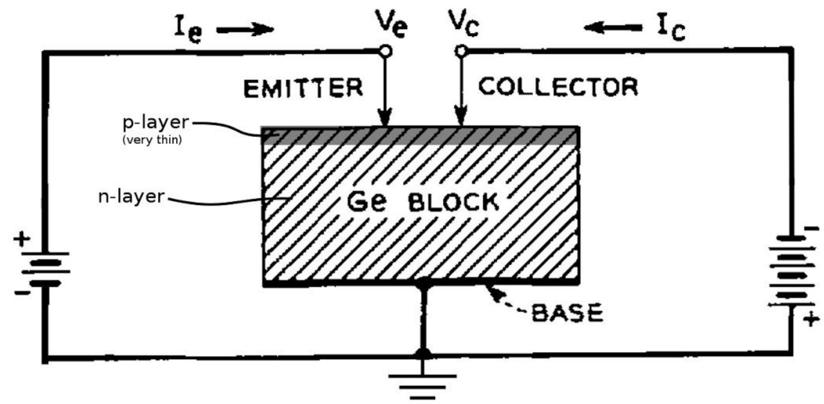
I had to figure out how to implement resistor-transistor logic using a point-contact transistor using saturated current. This was never done at the time because, as noted below, BJTs (bipolar junction transistors) soon replaced point-contact transistor. The operation of a BJT as a logic gate is much more familiar to a modern electrical engineer. In the footnotes I mentioned some of the differences between these two types of transistors. The reason I use a point-contact transistor is so that in the video I can describe the operation of the first transistor (a point-contact one) as an amplifier and then develop the essentials of transistors as logic gates without having to introduce another type of transistor.

For this document I have used as much data as possible to ensure that I don’t get things wrong. Specifically, figure 3 in Bardeen and Brattain, 1949.

Note: I have skipped some of the discussion of electrons and holes in semiconductors because it is my book *Eight Amazing Engineering Stories. p. 197-202.*

Point-Contact Transistor as an Amplifier

The first use of the transistor was as a current amplifier.[1] The circuit they used is:

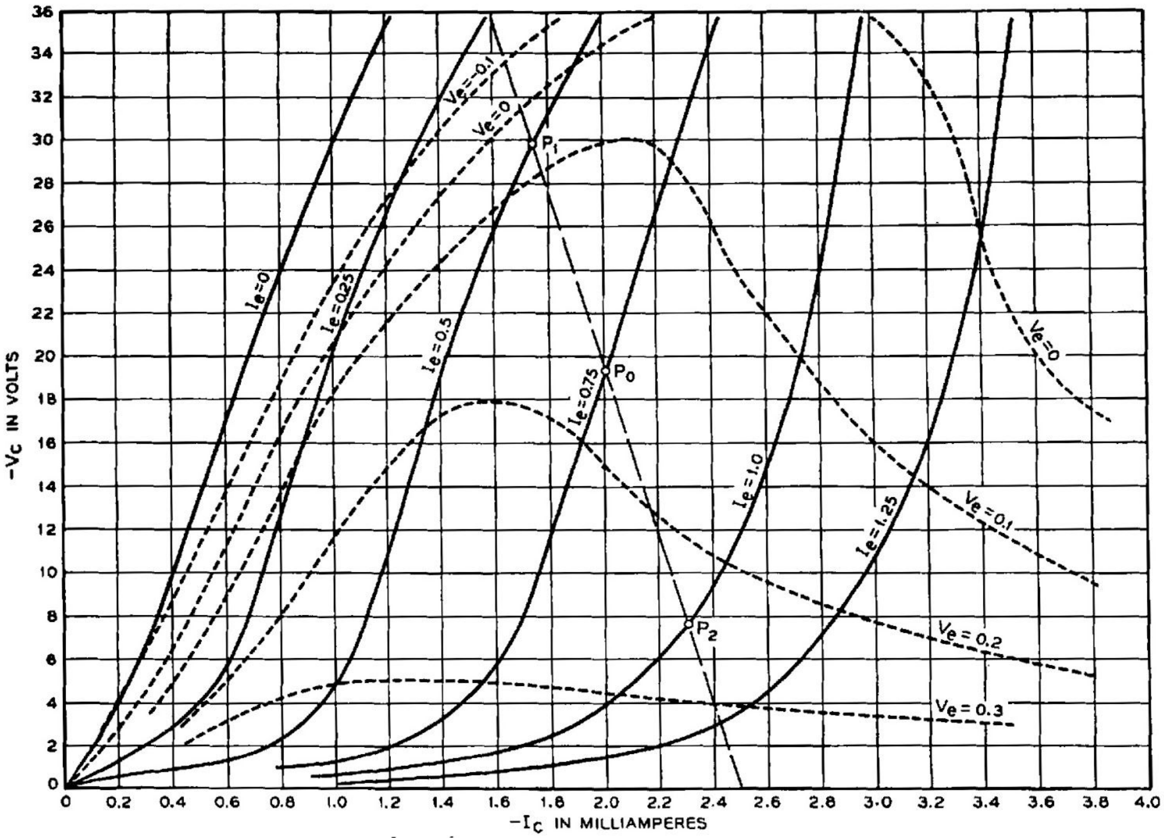


The germanium block is n-type with a very, very thin layer of p-type at the top. The emitter current i_E controls the collector current i_C. The important part here is that the current is amplified, or better the power is amplified. Here is how it does this: Look at the left side of the device as a pn junction diode: it is forward-biased and so current flows. If we imagine the right side also as a pn diode, we see it is reverse-biased — no current will flow. When we put them together the holes passing through the p-section will be "injected" into the collector side: this creates the collector current.

To understand its operation I use actual numbers from Bardeen and Brattain, 1949. As Bardeen and Brattain note, of the four variables (emitter current, emitter voltage, collector current, collector voltage)

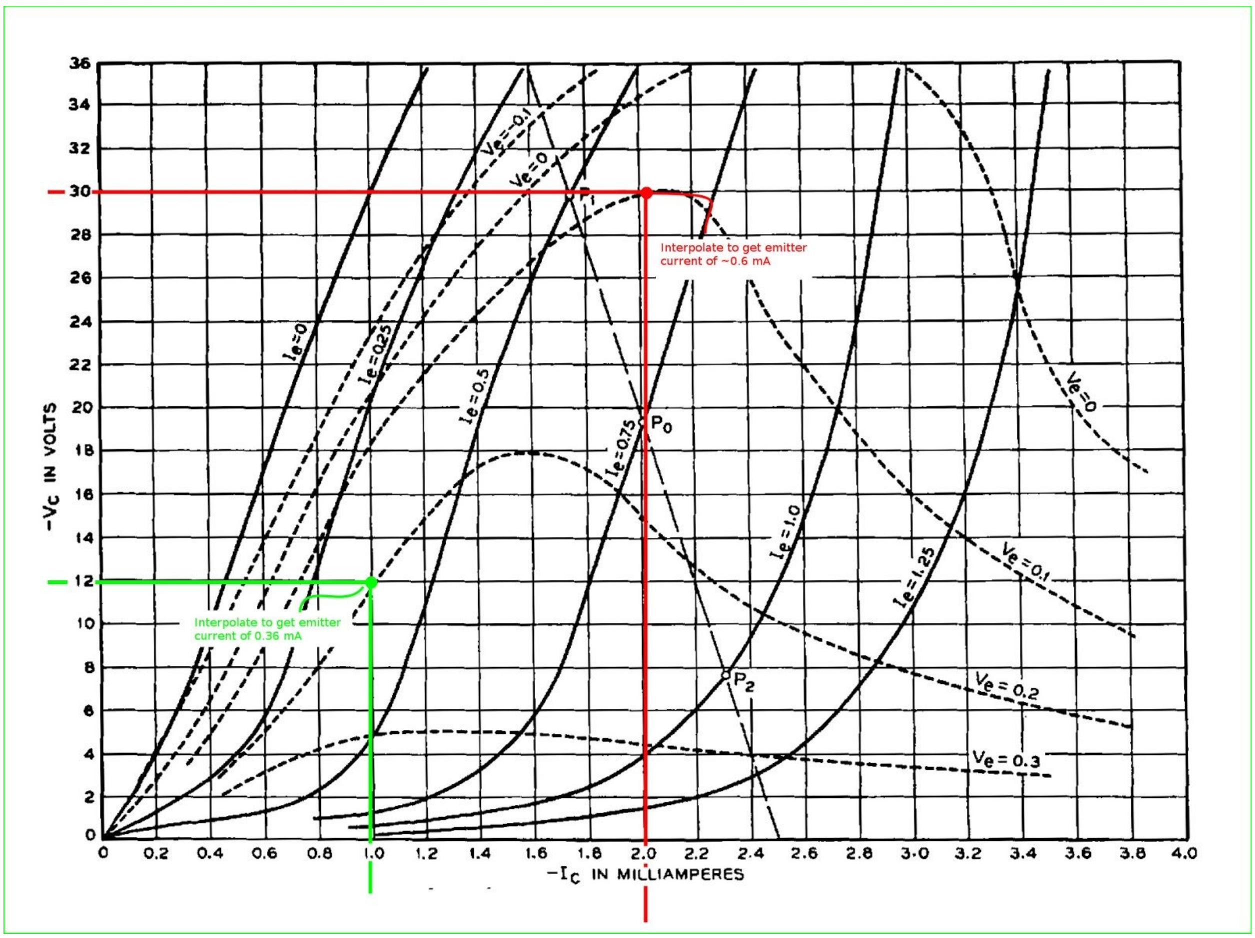
[1] Note that in some configurations, in contrast to a point-contact-transistor, a BJT does *not* amplify current, it amplifies power — most modern electrical engineers are puzzled by the current amplification in a point-contact transistor. For a clear description of a common-base, non-current amplifying BJT see pages 25-26 of Amos. Note also something else that can throw off a modern electrical engineering: This is a point-contact transistor with a common base, when BJTs are used they often have a common emitter and so the details are slightly different.

only two are independent — pick two and the other two are determined.

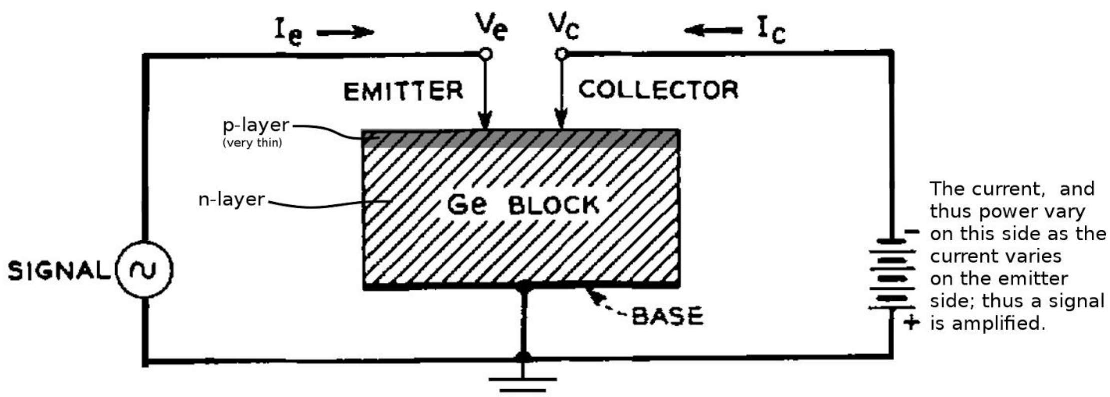


For example, if the battery for the collector were -12 volts and that of the emitter were +0.2 volts then we would have currents of i_E = 0.36 mA and i_C= 1 mA. (Where I got these numbers is shown in green in the graph below.) This is an amplification of the current of 1/0.36 = 2.78. Note, that the power is much, much greater. Power is *Vi* so the power on the left is 0.36 mA*0.2 volt = 0.72 mW and on the right 1 mA*12 volt = 12 mW. This would be a power increase of 16.7. Or expressed as decibels 10*log(16.7) = 12 dB.

Or, to get a great amplification use a V_E = 0.2 volts, V_C = 30 volts, which yields i_E = ~0.6 mA and i_C= 2 mA. (Where I got these numbers is shown in red in the graph below.) This is a current amplification of 2/0.6 = 3.33; power amplification of (30*2)/(0.2*0.6) = 500 or 27 dB.



The importance of this transistor as an amplifier occurs if we change the battery on left (emitter) side with a signal (a sound, for example) — a voltage that varies with time — while keeping the voltage on the collection side constant. The emitter current varies with time because the emitter voltage varies with time, and so the amplified current (the collector) varies with time, and thus the power of the input signal is amplified. The first uses of these transistors was in small radios and in hearing aids.



Using a point contact transistor as a switch

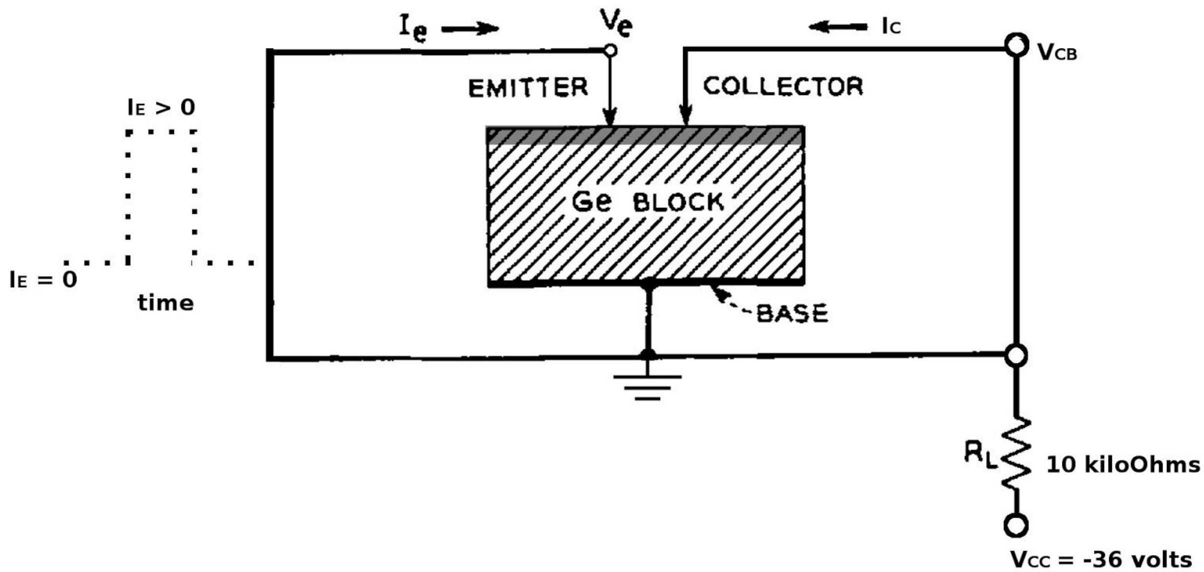
To use the transistor as a switch we set the emitter current with a pulse "by an unspecified current source."[2] The idea is this: The transistor is "off" if no emitter current flows, and it is turned on if a current pulse (which is a "true" or "1" in our binary logic) is large enough to drive the transistor to saturation (explained in more detail below), i.e., the point where the collector current is at a maximum, and where it can not go higher even if the emitter current is higher.

As we will see, saturation is important because: a) We always have the same output from the transistor when it is "on" (not different values of current) and b) if we have multiple input they don't increase the output current, which is useful for an OR gate, for example. (See below.)

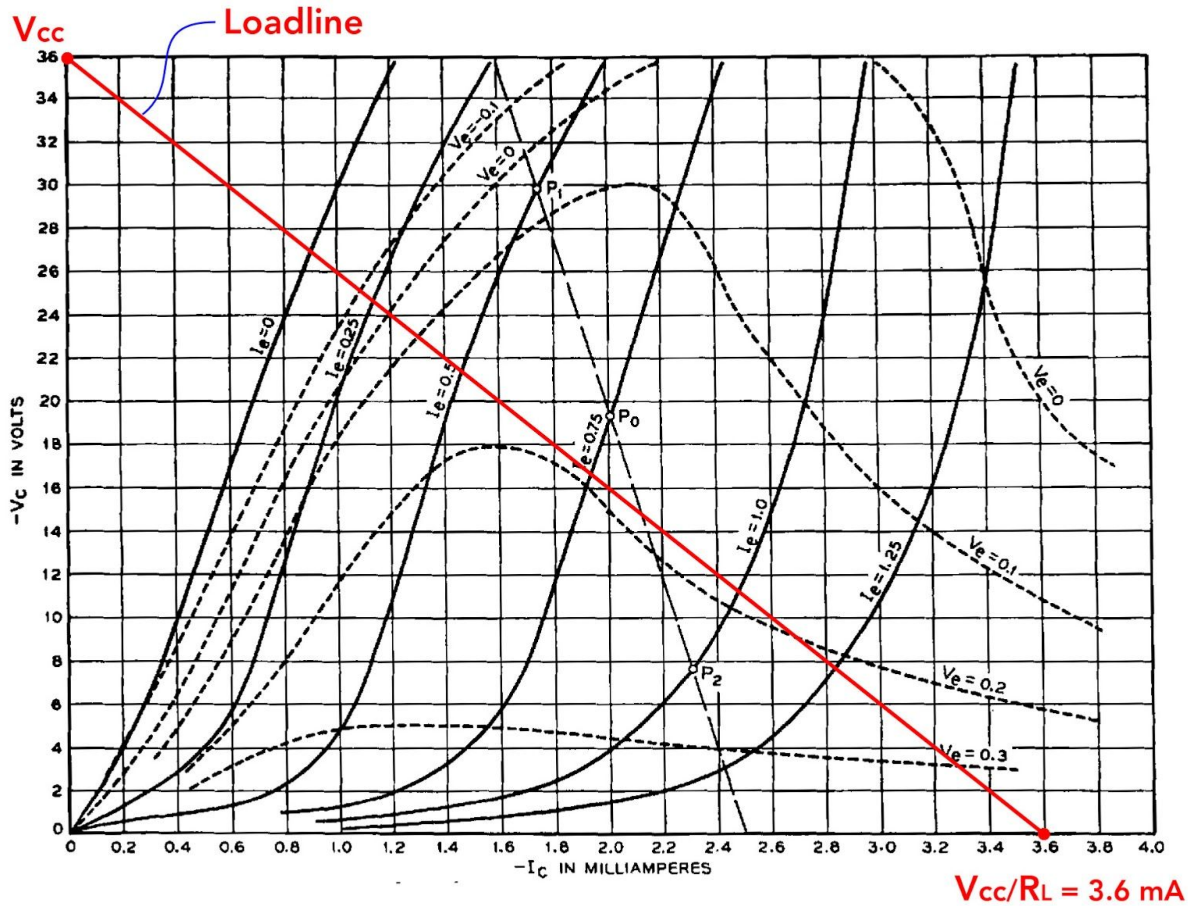
Note that this ability to use an input into the emitter to trigger from one stable state to another ("off" to "on") is not possible with a *single* vacuum tube. (See Braun, p. 129.)

[2] This is from Warner, 1991. This book on pages 518 and 519 greatly helped me to understand how a transistor is used as a switch. The circuit on page 519 (which is similar to the circuit here) was the key. I was puzzled about how the voltage on the collector side could be changed — I knew that at no point was the that voltage being changed by a signal to the collector side, it had to be driven by the emitter pulse — which indeed it is as described above. In general, this is a very good, very detailed book that takes one slowly through all aspects of transistors.

To understand this operations consider this circuit:



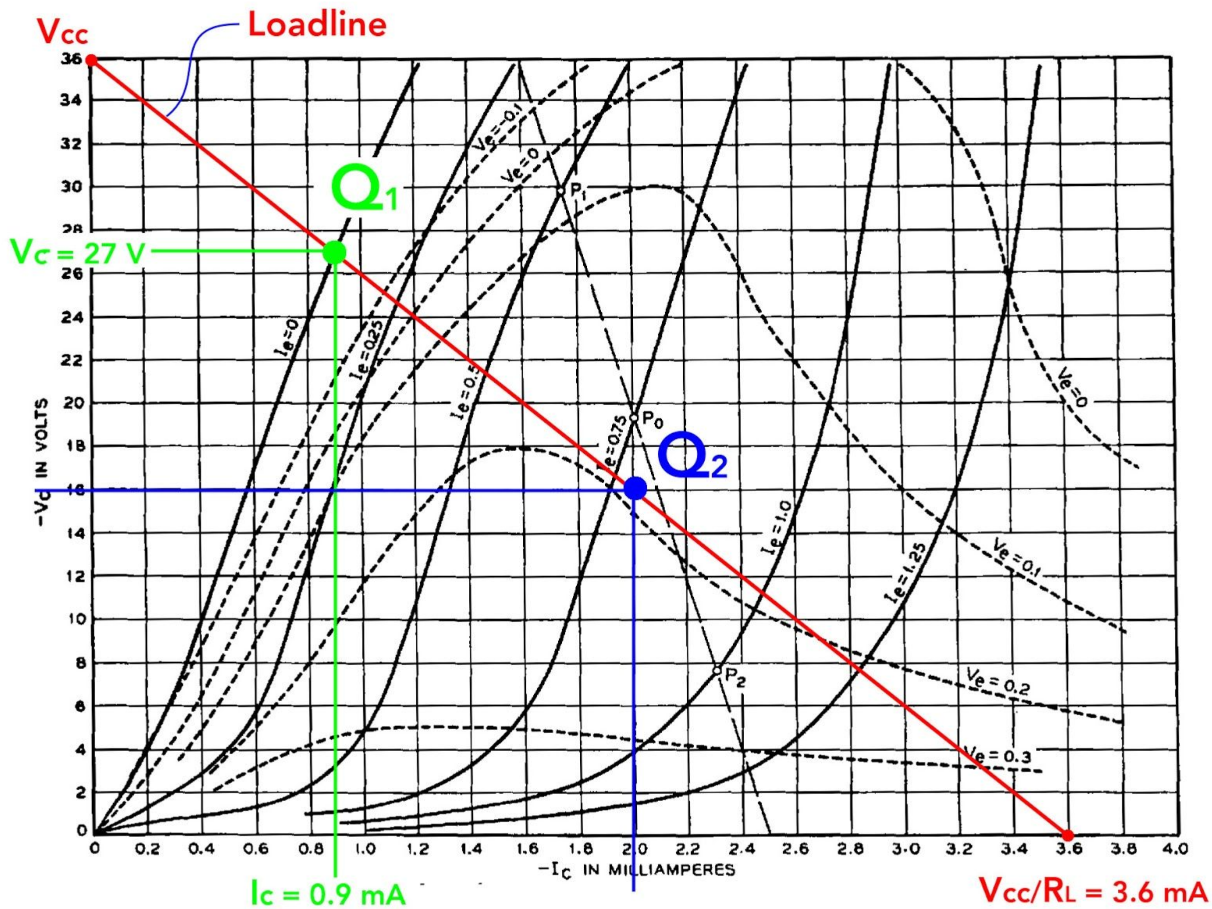
The first thing we need to do with this circuit is add the load line to the iV data from Bardeen and Brattain. The key idea is this: when linear and non-linear elements exist in a circuit you need to be sure that Kirchhoff's voltage law is obeyed: Kirchhoff's current law (1st Law) states that current flowing into a node (or a junction) must be equal to current flowing out of it. To do this we draw a “load line” on the diagram. It represents the constraint put on the voltage and current in the nonlinear device by the external circuit. The load line, usually a straight line (ohm’s law), represents the response of the linear part of the circuit, connected to the nonlinear device in question. The points where the characteristic curve and the load line intersect are the possible operating point(s) of the circuit; at these points the current and voltage parameters of both parts of the circuit match. On the current axis the line cross at V_{max}/R and on the voltage line is crosses at V_{max}. Here V_{max} is the voltage V_{cc}. The load line for the conditions in the circuit above is in red:



The device can operate anywhere on this line. Shown below are operating points (for example) Q_1 and Q_2. At Q_1 the emitted current is zero, the voltage V_E is ~ -0.25 Volts, I_c = 0.9 mA, and V_{CB} = -27 volts. Notice what the load line does: A current of 0.9 mA flows through the collector side, this must flow through the resistor R_L. Thus, the voltage drop over the resistor is 0.9 mA*10 kOhm = 9 volts. The drop from V_{cc} is -36 + 9 = -27: *exactly* the correct voltage for V_{CB}.[3]

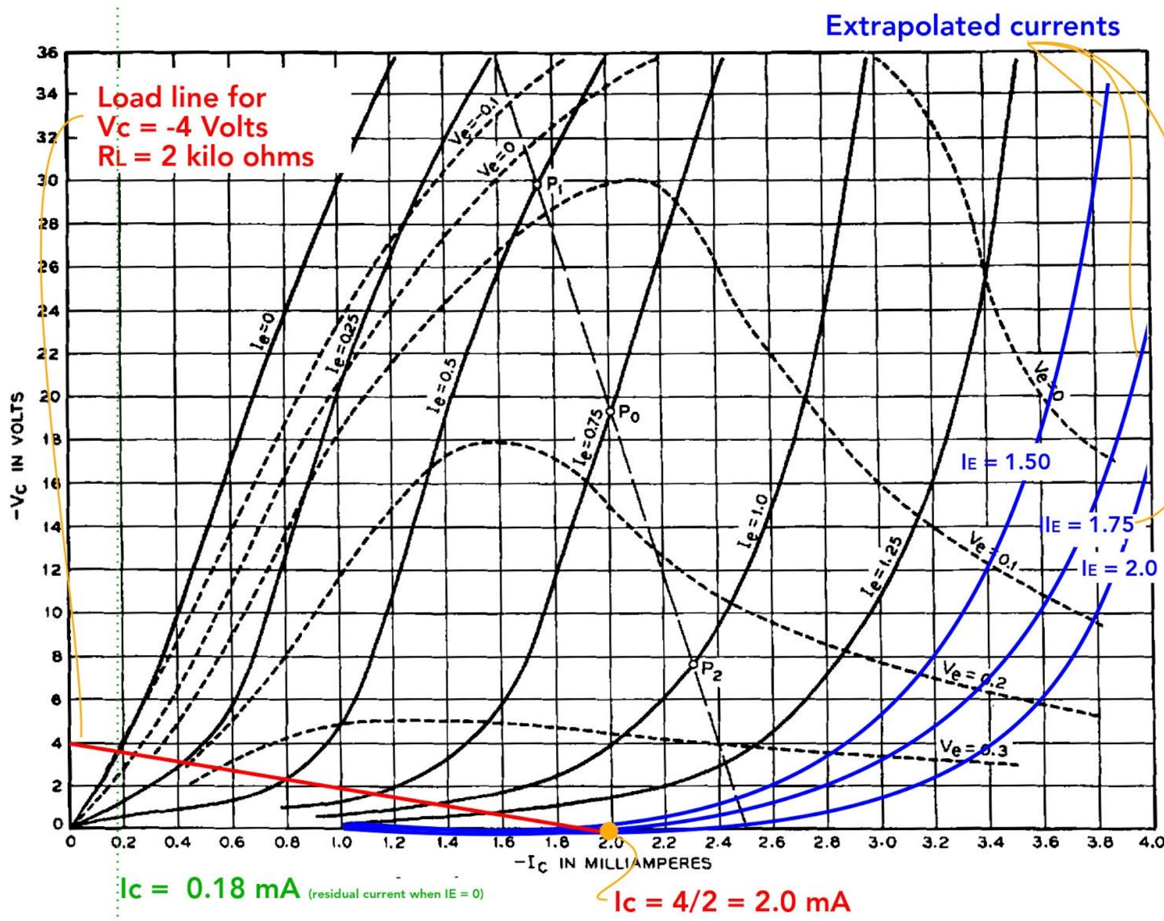
[3] This behavior explains something I was missing (and which gets us ahead of ourselves) is that the voltage on the output side of the transistor varied when the transistor was used as a switch: It was forward biased when the transistor was on, but negative biased when it was off. I could not understand what controlled the voltage, now I see it is done *by the circuit!* As I learned later: “Using a property chosen resistive load together with a voltage or a current source, one can obtain bistable operation of these devices [point-contact transistor, unijunction transistor, the four-layer switch and the silicon-controlled rectifier (SCR)].” p. 393 of Pederson and Mayaram *Analog Integrated Circuits for Communication: Principles, Simulation and Design* (Spring, 1991).

At Q2 you can verify this again with: I_E = ~0.756 mA, V_E = ~ 0.19 V, I_C = 2.0 mA, and V_{CB} = 16 V. (Note: The I_E and V_E are interporlated — perhaps badly!)



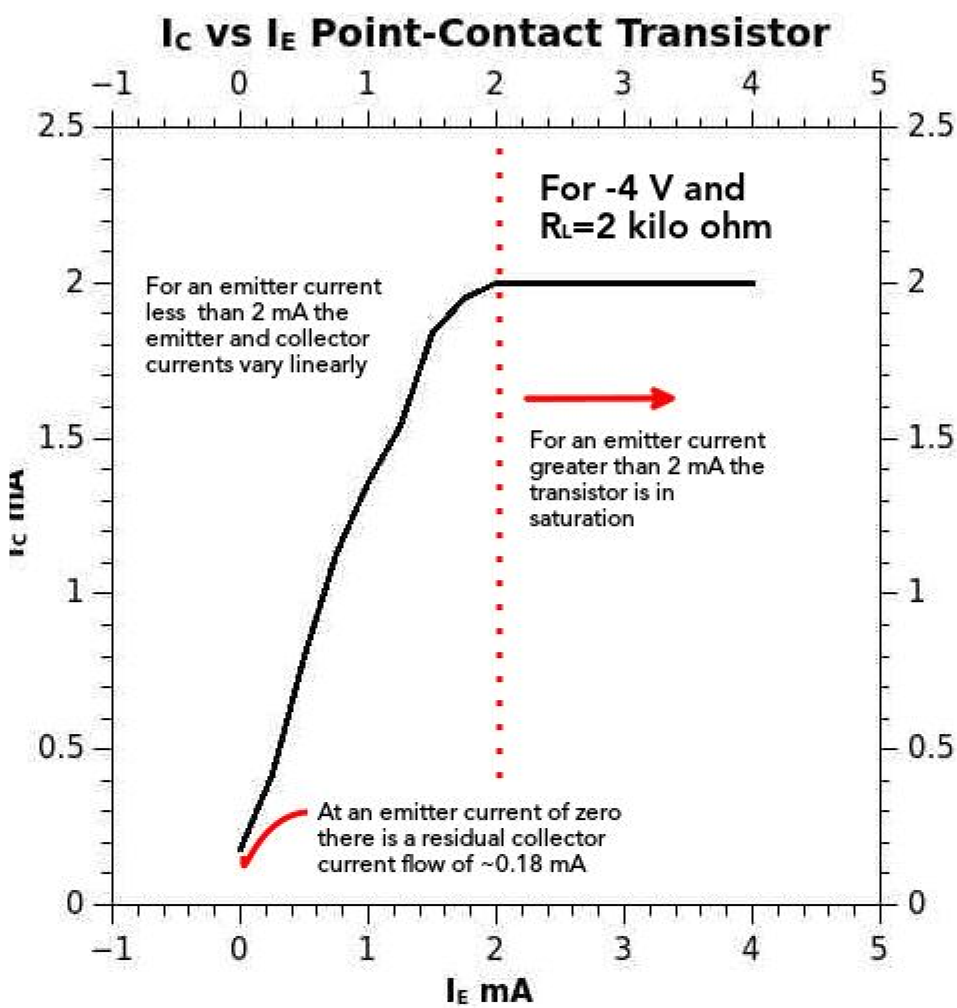
To use this as a switch we would love for this to happen: When zero emitter current flows, then there is no current flowing on the collectorside (this is “off”, “false”, or “0” — they are all the same) and then when we send a current pulse of a “particular” size (and we want to send the same size every time) we get a standard sized pulse from the collector side. We cannot quite do this, but we can come close.

To see this look at a circuit were V_{CC} is -4 volts and the resistor is 2 kΩ. I've drawn this load line on the graph below. I have also added to Bardeen and Brattain's data additional emitter currents.[4]



[4] These are extrapolations because little saturation data exists for a point-contact transistor; they were soon replaced by BJT. There is some data in Bassett and Tillman (1953) and in Meacham and Michaels (1950).

If we plot i_E vs i_C (as read off the load line moving left to right) we would see this: a) when $i_E = 0$ we have a residual collector current of 0.18 mA, b) as the emitter current increases, the collector current increases linearly, albeit amplified — this is the standard operation of the transistor as an amplifier, and c) as the emitter current gets higher the collection current plateaus at 2.0 mA when the emitter current is above 2.0 mA: it will not rise higher no matter how much larger the emitter current gets — this is called *saturation*.



This explains how we use the transistor as a switch. We use the emitter current of zero to mean part of a syllogism is false — we'd like zero output, but we don't get that, so we expect (or, better set our electronics) so that a current of 0.18 mA is a zero. When a statement is “true” (“on” or “1”) we send a current pulse of at least 2 mA into the transistor — note this is the minimum current that will put the transistor into saturation.

What we have now is a switch or a “logic gate.” When it gets a “false” input, that is, no emitter current, it reports “false”, that is, the switch is “open” — there is only a minimal current flow. When we send it a “true” — a current pulse of at least 2 mA — then we get a pulse of 2 mA, the saturation current, which corresponds to true. This means that we can construct logic from the point-contact transistor operated as a switch.

Logic gates

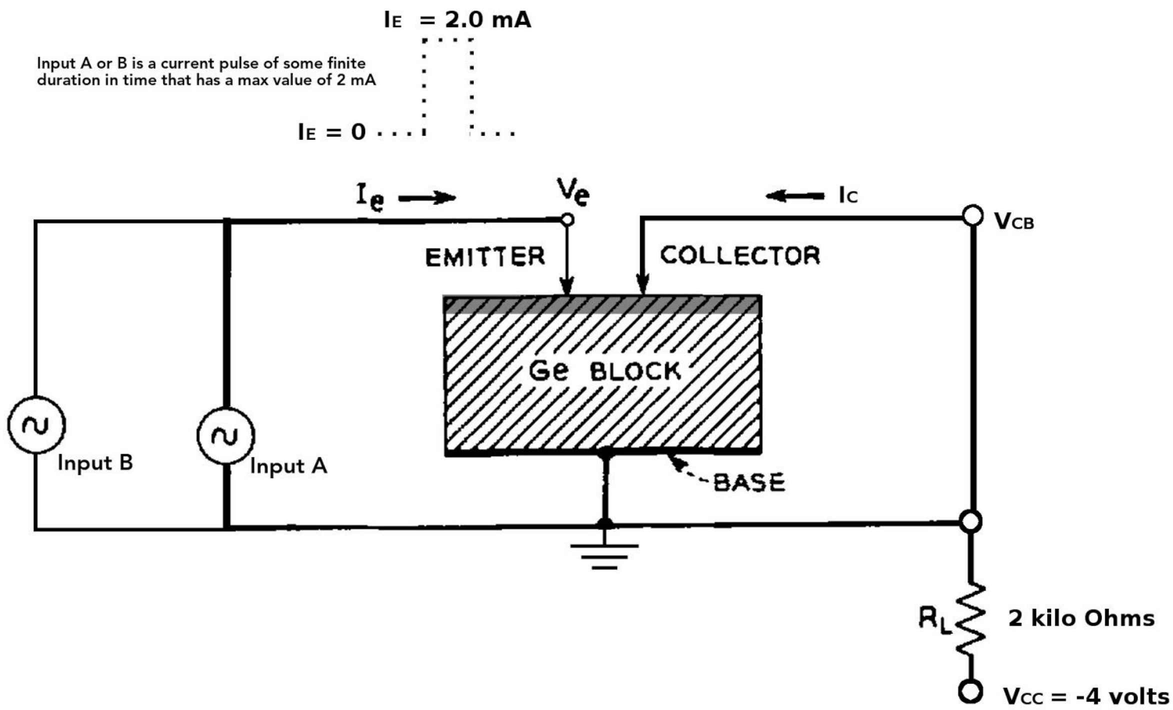
OR gate

For an OR gate we expect this kind of truth table:

A OR B = C

input A	**input B**	**output C**
false	false	false
true	false	true
false	true	true
true	true	ture

To create an OR gate from the a point-contact transistor we use the following circuit, where the inputs are in parallel:



For this circuit we have two inputs labeled “A” and “B” here. We can send a current pulse to a) neither, b) either, or c) both. Look what happens:

If we send no current pulses through “A” or “B” then only residual current flows in the collector side, which indicates “false.”

If we send a current impulse of 2 mA through “A”, and none to “B” the we get a 2.0 mA output, which indicates the statement is true.

If we send no current impulse through “A”, and a current pulse of 2 mA mA through “B”, we get a 2.0 mA output, which indicates the statement is true.

If we send a 2.0 mA current impulse through both “A” and “B” we get an output of 2.0 mA. Note: This is where saturation is important — even though the current going in is double that of the two previous operations, the output current is the same because we are at saturation. **The non-linearity of the transistor is key here!**

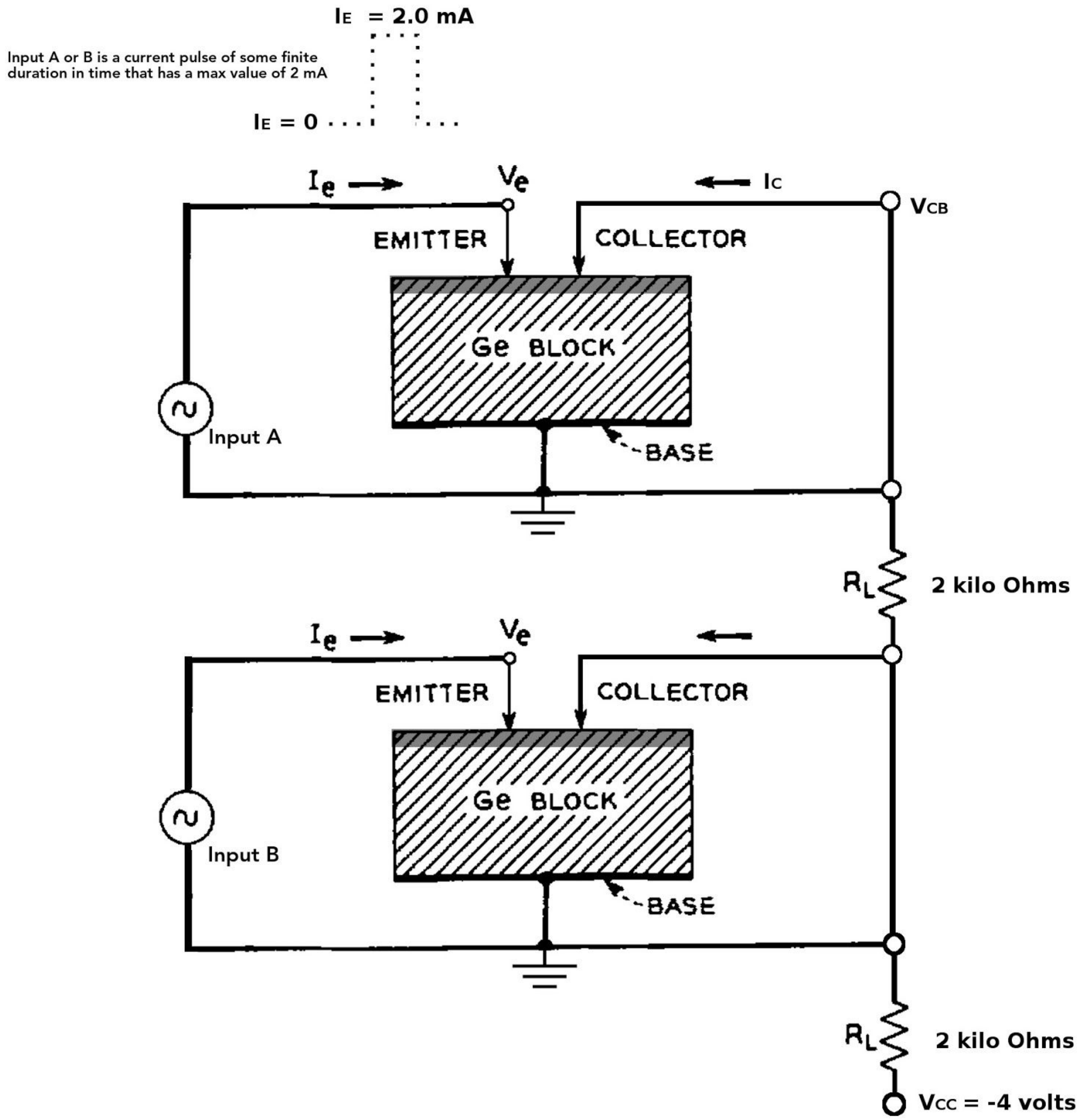
AND gate

The logical AND has this true table

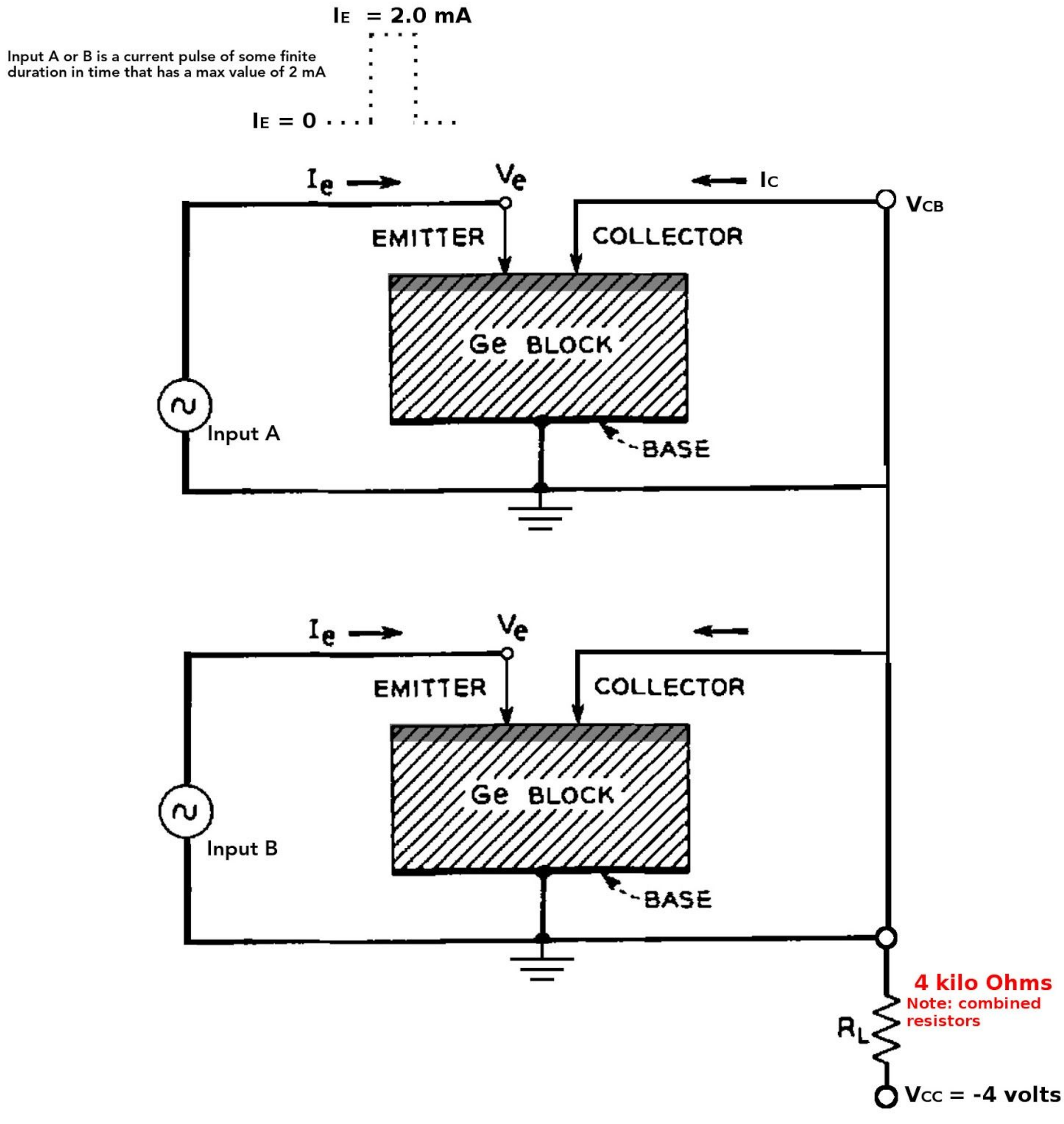
A AND B = C

input A	**input B**	**output C**
false	false	false
true	false	false
false	true	false
true	true	ture

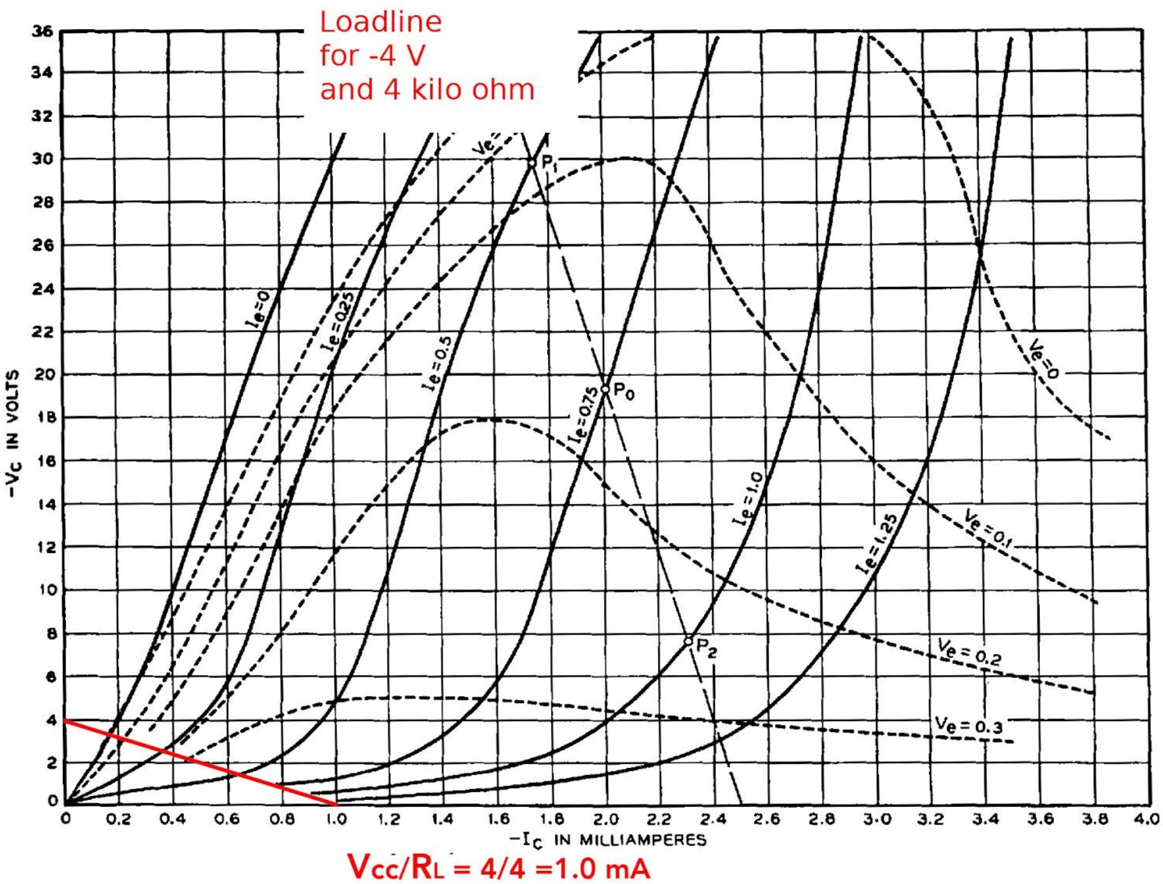
We can implement that with this circuit created by placing two transistor switches in series:



We can replace this circuit with one where the two transistors are combined into one of double the size, as shown below.



Notice that the load line for this combined transistor has changed: The load line now crosses the x-axis at 1.0 mA (half of the saturation current previously.) When either of the transistors saturates (i.e., an impulse of 2 mA) then in the collection 1.0 mA will flow. Note, too, that that residual current is now still about 0.18 mA.



Consider these four cases:

Input A and Input B are zero. Then 0.18 mA flows from each transistor for a total of 0.36 mA. This is below our threshold for “true” of 2.0 mA. We get a “false” signal from our logic gate.

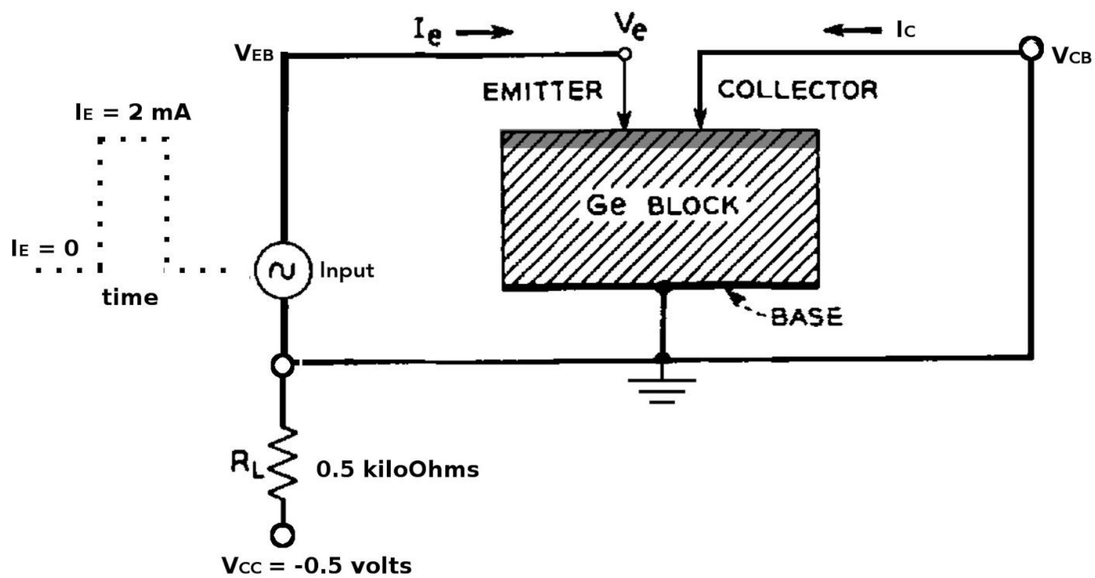
If Input A has a pulse of 2 mA and input B has no current pulse then 1.0 mA flow from the first transistor, and 0.18 mA (the residual current) from the second for a total of 1.18 mA — still below our 2.0 mA threshold. We get a “false” signal from our logic gate.

If Input A has no pulse and input B has an impulse of 2 mA then 0.18 mA (the residual current) flows in the first transistor, and 1.0 mA from the second for a total of 1.18 mA — still below our 2.0 mA threshold. We get a “false” signal from our logic gate.

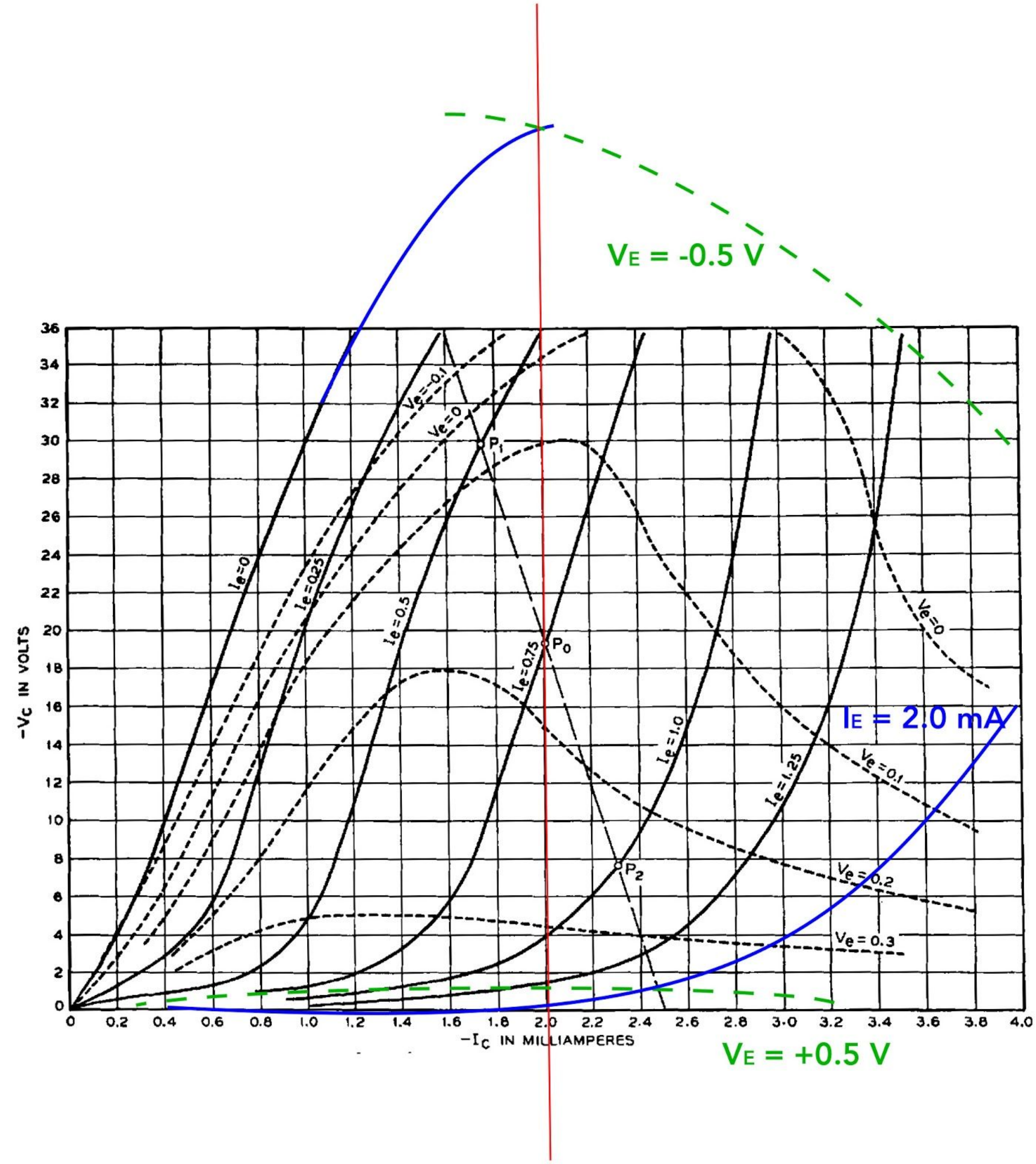
If both input A and B received a current impulse of 2 mA then we 1.0 mA flows in both transistor for a total of 2.0 mA — this meets our threshold and we get a “true” signal from this logic gate.

<u>NOT gate (inverter)</u>

An inverter (or NOT gate) take a “null” signal (i_E = 0) and generates a “on” signal in the collector (2.0 mA), or take a current pulse (i_E = 2 mA) and generate no current in the collector. To design the circuit below I had to do gross extrapolation on the Vi plot of Bardeen and Brattain — again, little saturation data is available.



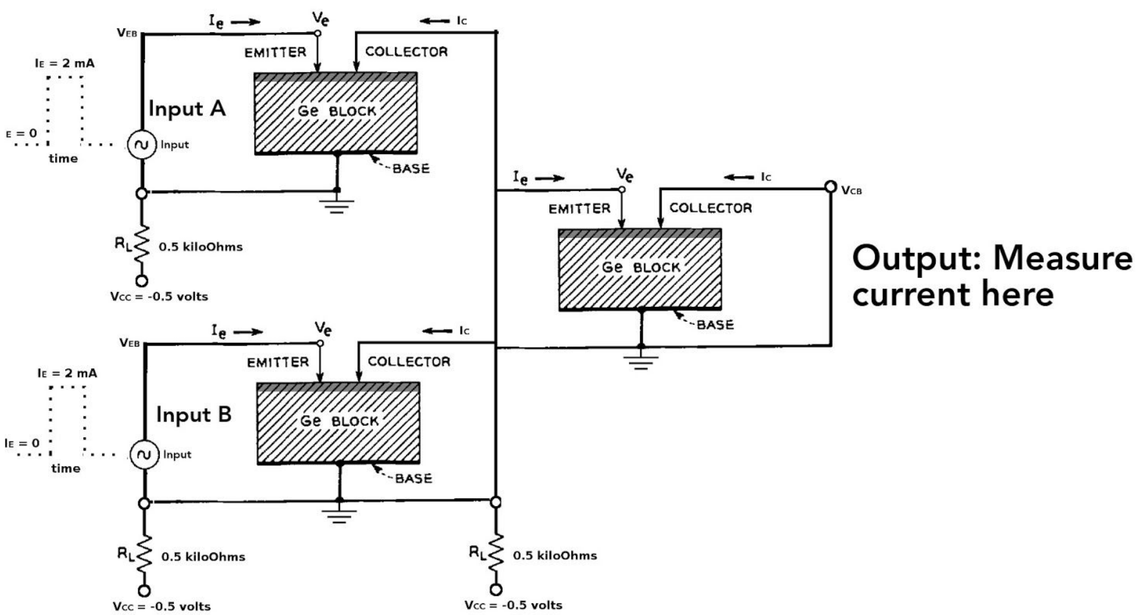
I added below values for V_E = -0.5 V, V_E = +0.5 V, and the current I_E = 2.0 mA.



Here's how this circuit operates. When there is no current in the emitter side (i_E = 0) the voltage at V_{EB} = -0.5. Note from the graph (the red vertical line) that for these conditions a current of 2 mA flows on the collector side. Thus the collector is "on." When a current pulse of 2 mA is fed to emitter side there is a voltage drop of 0.5*2 = 1 volt across the resistor and so the voltage at V_{EB} = +0.5. Under these conditions only a residual current flows in the collection and so it is "off."

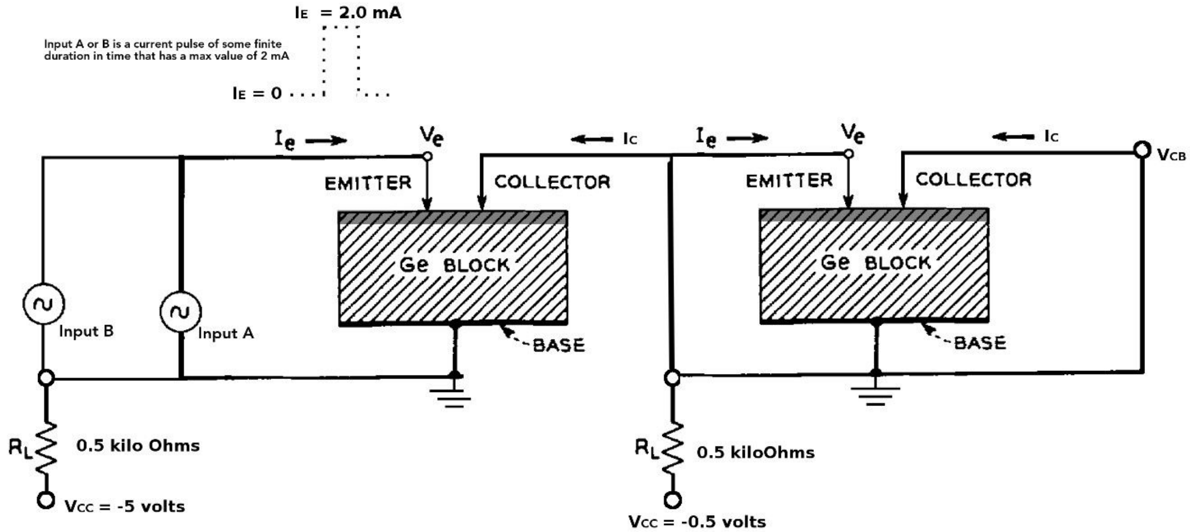
A better AND gate

Likely a better AND gate is created by combining two inverters. In the gate below you get only the residual if either of the two inputs are low ("false"), if they are both high ("true") you get the saturation current. I'm guessing this is better than the AND above which in one of its "off" modes has a current running that is half the saturation current.



Alternate OR gate

An OR gate can be made by putting two inverters in series, where the first inverter has inputs in parallel.



Why point contact transistor were not used to implement logic

Point contact transistors were difficult to manufacture reliably and were soon replaced by BJT. For this reason they were never used for logic gates. In addition, point-contact transistors work in a saturated current mode: this is slow — it takes time for the holes and electrons to decay once the transistor is switched on.[5] In fact, the paper listed in the sources by Bassett and Tillman focuses on the delayed responses once the current is switched off. BJT were used in a voltage mode that solved overcame many of these problems.

When point-contacts transistor were used in a computer (e.g., the TRIAD designed by Jean Fellker in the early 1950s), the logic was done by diodes and the transistors just did amplification.

[5] See Braun on p. 129 "an obstacle to high speed operation of point-contact and junction transistor arises from delays in response due to the phenomena of saturation and hole storage." In addition, switching using saturated current doesn't work very well: BJTs work in a voltage mode instead. As Braun notes, on page 147, "All the transistor switch circuits described thus far operate in what may be termed a voltage mode . . . [but one could use] circuits operating in a so-called current mode, wherein the current from an essentially constant current is switched. Circuits of this type may be designed for either saturated or non-saturated operation. The discussion following will be confined to nonsaturated current switching systems, these being capable of higher speed operation." He seems to imply that using saturated current operation is too slow.

Sources

Amos, S.W. *Principles of Transistor Circuits* (Newnes, Eighth Edition).

Bassett, H.G., and J.R. Tillman "Some transient properties of transistors," *British Journal of Applied Physics* **4** (April), p. 116 (1953).

Braun, Edward L. *Digital Computer Design: Logic, Circuitry, and Synthesis* (New York: Academic Press, 1963).

Bardeen, J. and W. H. Brattain "Physical Principles Involved in Transistor Action" *Bell System Technical Journal* 1949, vol. 28, no 02, p.239-277.

Mecham, L.A. and S.E. Michaels "Observations of the Rapid Withdrawal of Stored Holes from Germanium Transistors and Varistoras," *Physical Review* 7**8**, p. 175 (1950).

Warner, R. M. and B. L. Gung, *Semiconductor-Device Electronics* (Philadelphia: Holt, Rinehart and Winston, 1991).